



ALFRED P. SLOAN
FOUNDATION



The Landscape of Data Sharing and Computational Reproducibility for Social Research

Pew Research Center

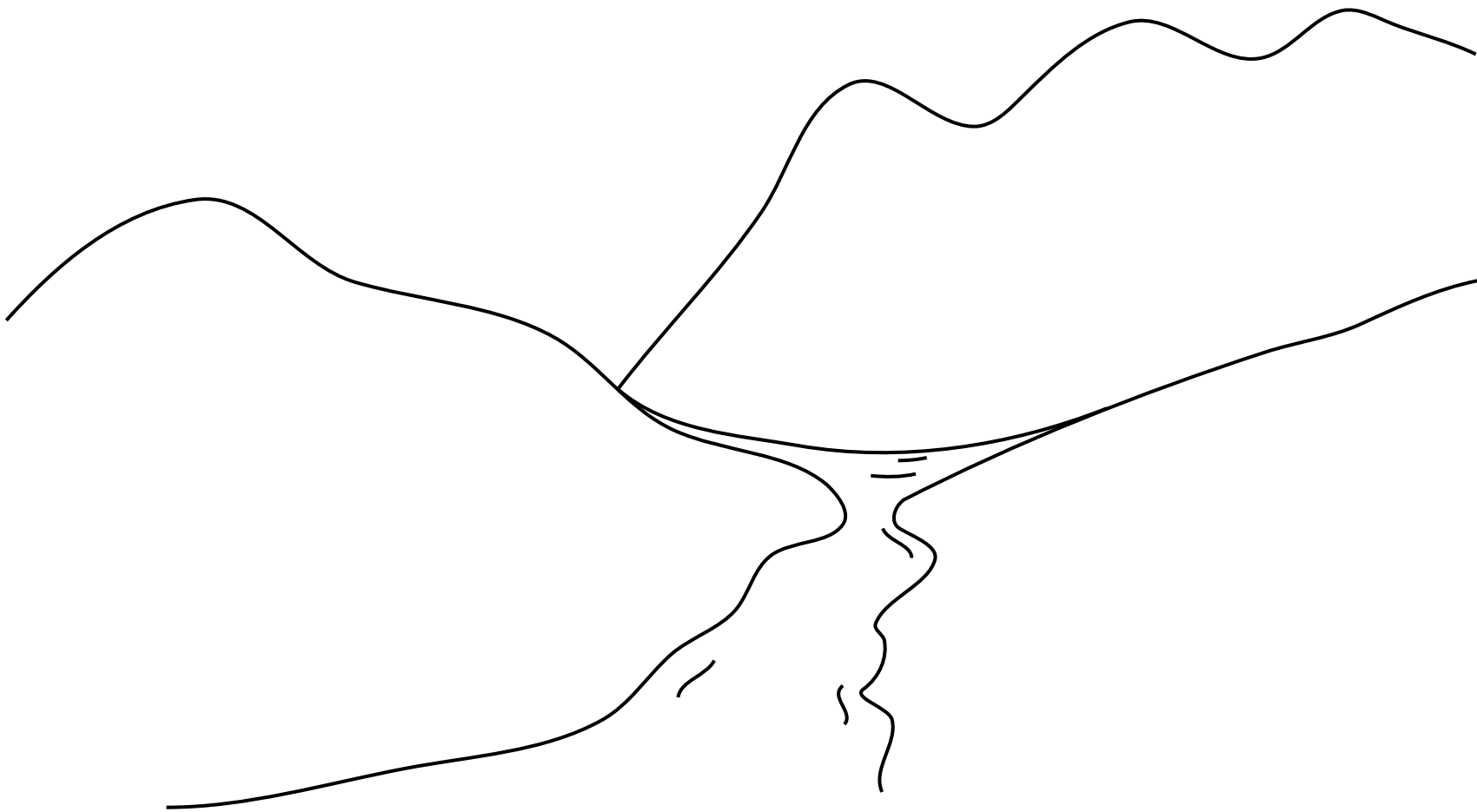
Feb 25, 2022

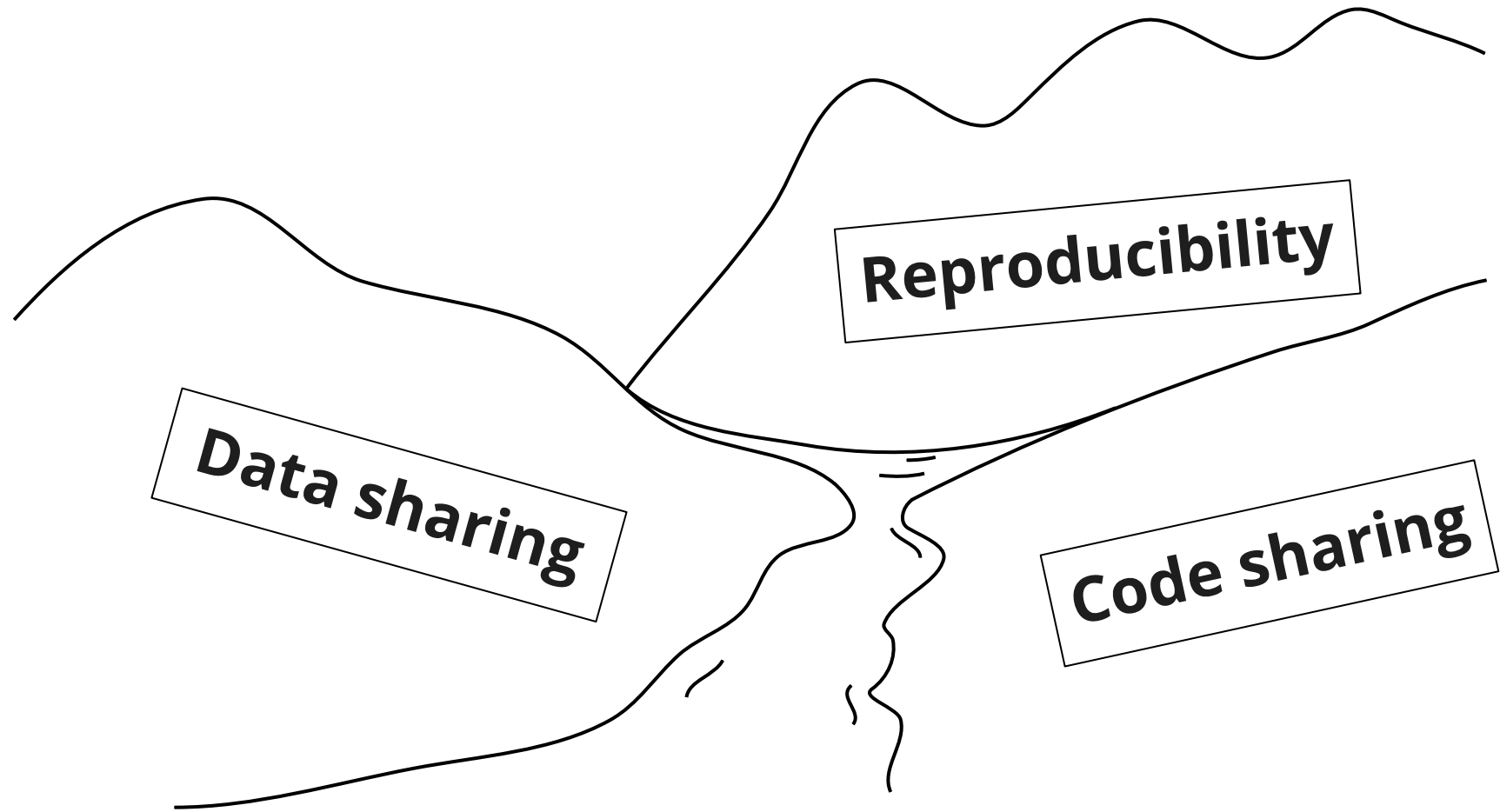
Ana Trisovic, Harvard University

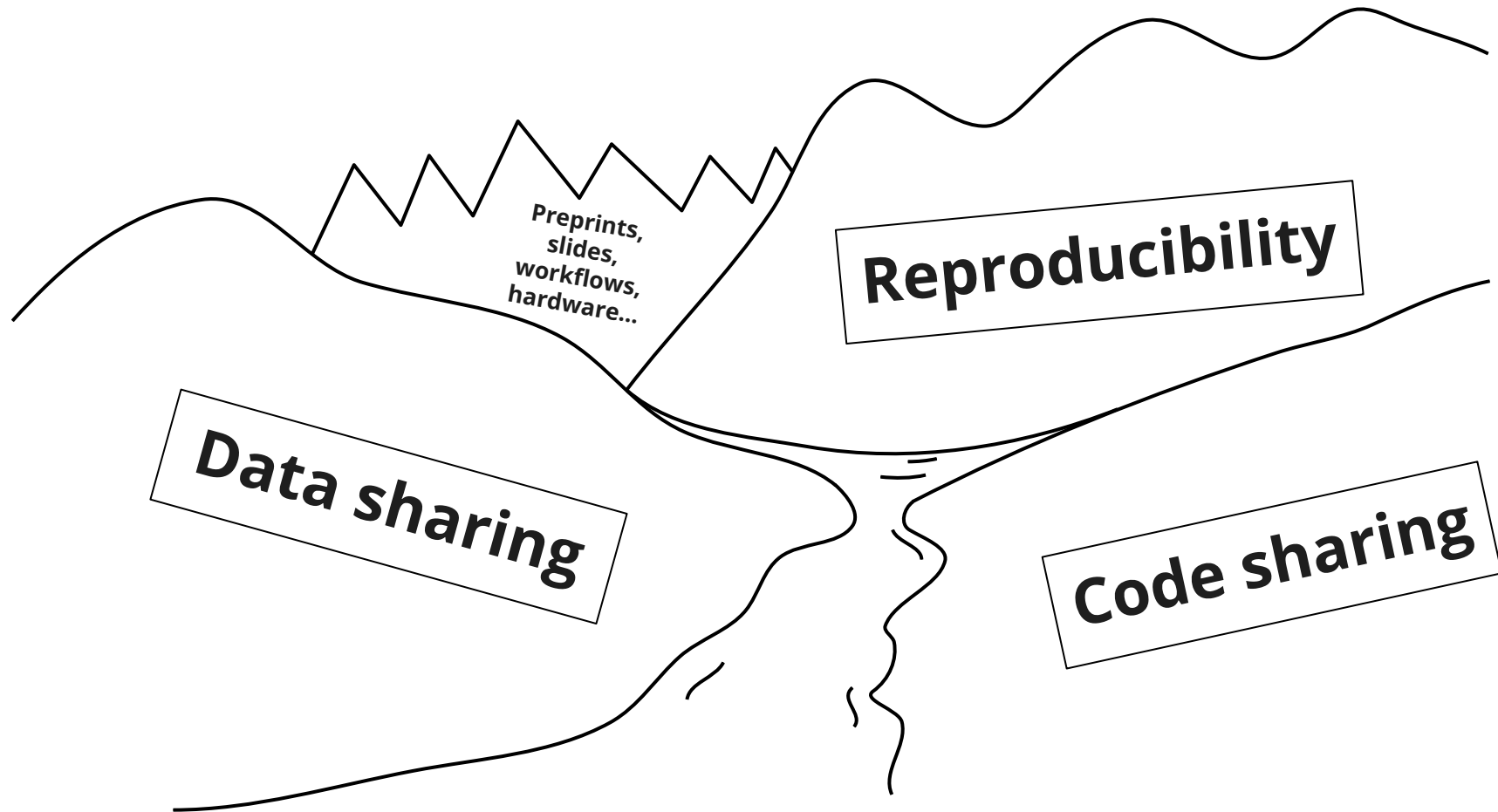
“An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the software, [data] ... and set of instructions which generated the figures. ”

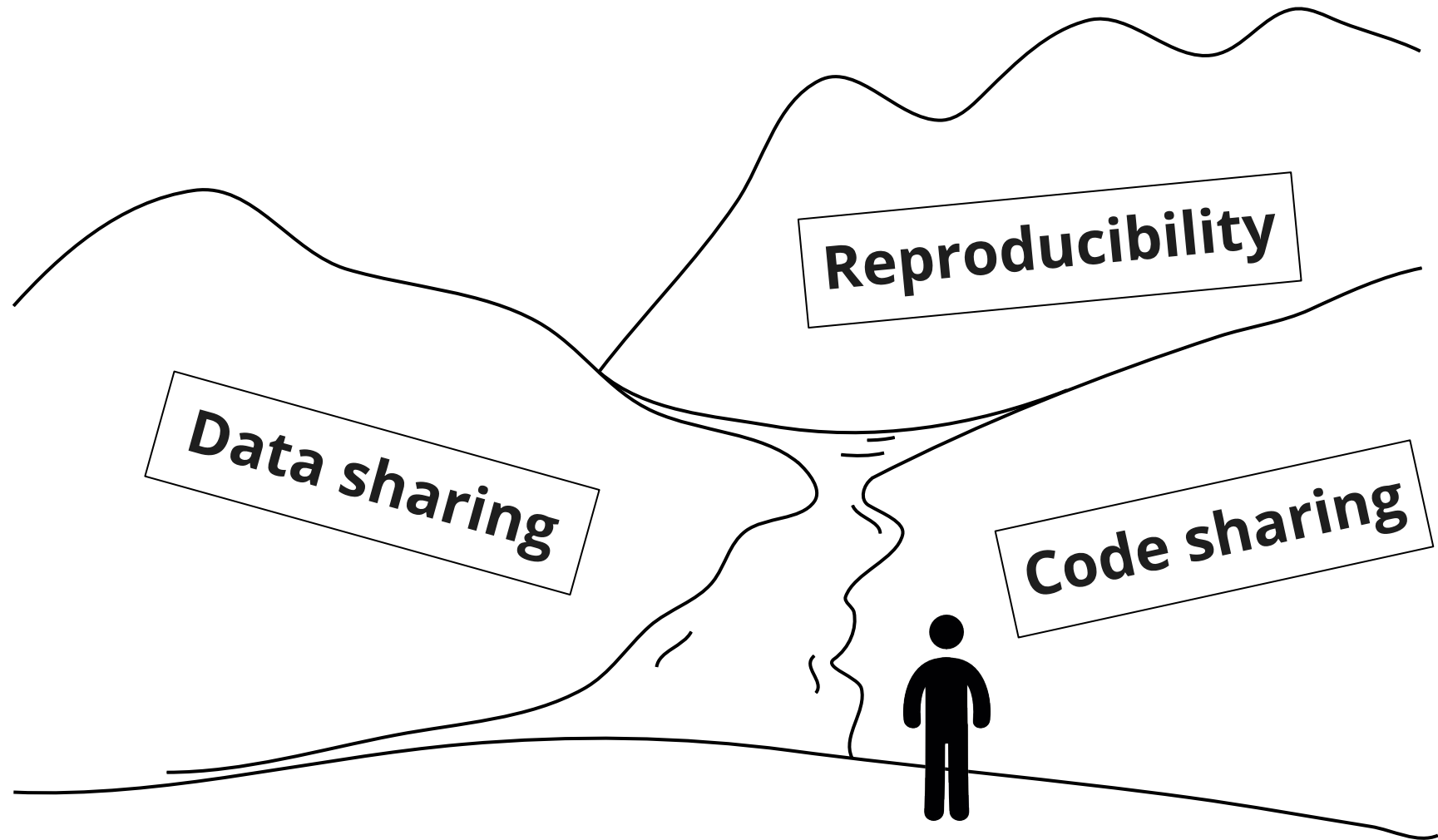
~ Prof Claerbout

Reproducibility: “obtaining consistent computational results using the same input data, steps, code, and conditions of analysis”





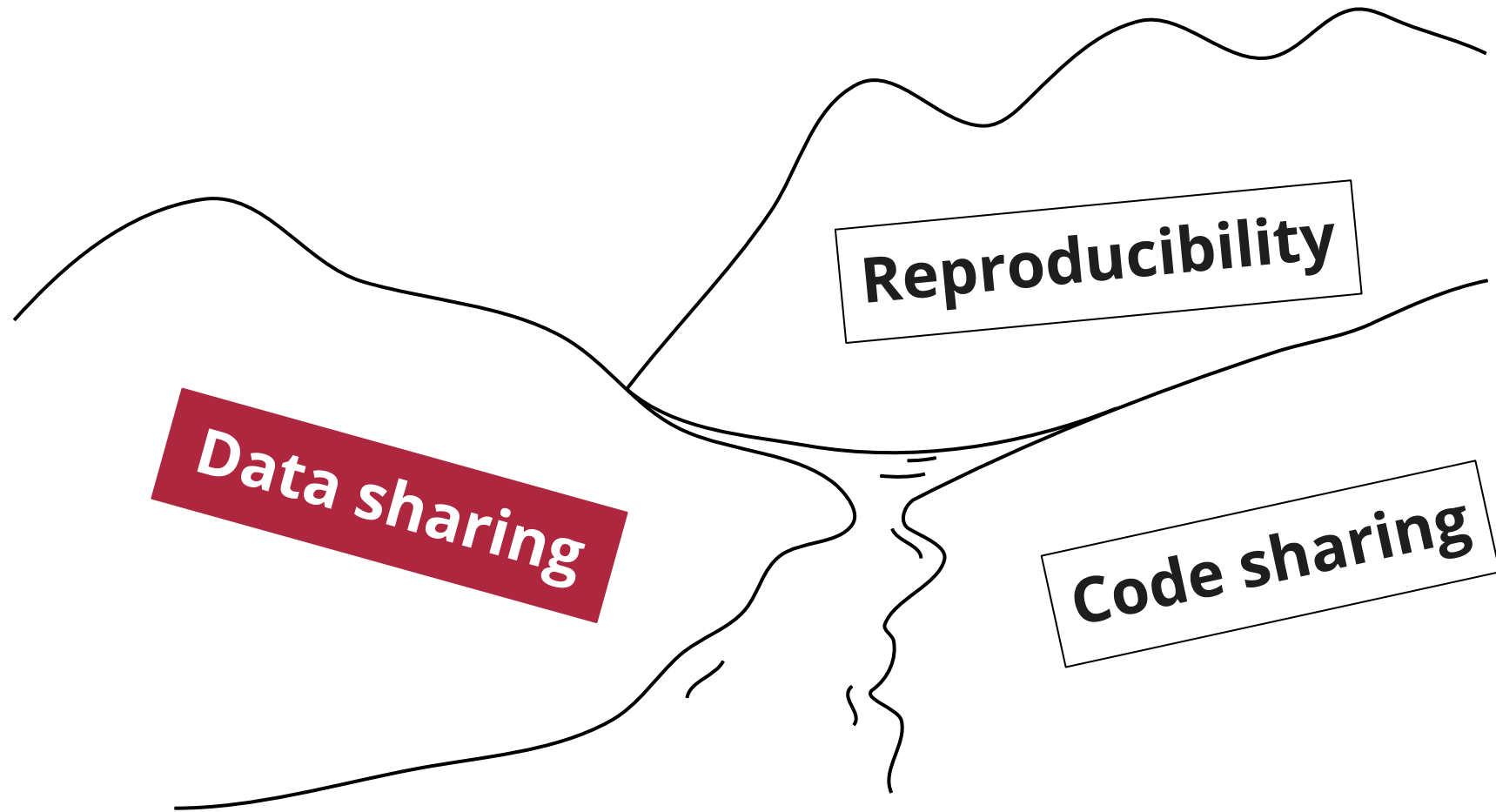




Data sharing

Reproducibility

Code sharing





- A free and open-source software platform to archive, share, and cite research data
 - Focus on data sharing and making data available

76 institutions around the globe run Dataverse installations as their official data repository



Findable
Accessible
Interoperable
Reusable

Findable	Describe data in metadata, assign DOI Metadata record is shared in data repository
Accessible	Accessible but not necessarily open Standard access protocol
Interoperable	File format open or proprietary Formal knowledge representation
Reusable	License and usage rights Data provenance

- ## Two approaches for data sharing:
- Via UI in the web browser
 - Via API (command line or Dataverse Software clients)

The screenshot displays the Harvard Dataverse web interface for creating a new dataset. At the top, the Harvard Dataverse logo is visible on the left, and navigation links for 'Add Data', 'Search', 'About', 'User Guide', and 'Support' are on the right. The main section is titled 'Host Dataverse' and contains a text input field with 'Harvard Dataverse' entered. Below this, a note states 'Changing the host dataverse will clear any fields you may have entered data into.' A warning message reads '*Asterisks indicate required fields'. The 'Citation Metadata' section is expanded, showing several required fields: 'Title' (text input), 'Author' (a sub-section with 'Name' and 'Affiliation' inputs), 'Contact' (a sub-section with 'Name' and 'E-mail' inputs), 'Description' (a large text area), 'Date' (text input with a YYYY-MM-DD format), 'Subject' (a dropdown menu), 'Keyword' (a sub-section with 'Term' and 'Vocabulary' inputs), and 'Related Publication' (a text area). Each sub-section has a '+' button to its right. The 'E-mail' field contains the address 'anatrisovic@fas.harvard.edu'. The 'Date' field is empty. The 'Subject' dropdown is set to 'Select...'. The 'Term' and 'Vocabulary' fields are empty. The 'Vocabulary URL' field has a placeholder 'Enter full URL, starting with http://'. The 'Related Publication' field is empty.

Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0



Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", <https://doi.org/10.7910/DVN/Z5BTCQ>, Harvard Dataserv, V2, UNF:6:YTyX+kjbsSZUNEND/3GGg== [fileUNF]

[Cite Dataserv](#) -[Learn about Data Citation Standards.](#)

Access Dataserv -

[Contact Owner](#) [Share](#)

Dataset Metrics

1,092 Downloads

Description

How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinions changed immediately and substantially when their party switched its policy position—even when the new position went against citizens' previously held views. These findings advance the current, largely experimental literature on partisan elite influence. (2020-03-26)

Subject

Social Sciences

Keyword

Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey

Related Publication Bisgaard, Martin, and Rune Slothuus. [date]. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. <http://ajps.org/>**Notes**

This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the [Center for Open Science](#).

[Files](#) [Metadata](#) [Terms](#) [Versions](#)

Search this dataset...

Filter by

File Type: All - Access: All -

Sort

1 to 10 of 25 Files

[build_data.R](#)
R Syntax - 12.1 KB
Published Jun 29, 2020
56 Downloads
MD5: a94...597



[codebook ess.pdf](#)
Adobe PDF - 508.8 KB
Published Jun 29, 2020
46 Downloads



Dataset version

Unique DOI

Mandatory citation-level metadata

Center for Open Science Badges

Code, documentation and other files

American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) [ajps.org](#)

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse >

Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0

Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", <https://doi.org/10.7910/DVN/Z5BTCQ>, Harvard Dataverse, V2, UNK:YTyX+kjbsSZUNEND/3GGg== [fileUNF]

[Cite Dataset -](#) [Learn about Data Citation Standards.](#)

[Access Dataset -](#)
 Contact Owner | Share

Dataset Metrics [↕](#)
 1,092 Downloads [↕](#)

Description [↕](#)
 How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinion immediately and substantially when their party switched its policy position—even if their position went against citizens' previously held views. These findings advance the experimental literature on partisan elite influence. (2020-03-26)

Subject [↕](#)
 Social Sciences

Keyword [↕](#)
 Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion survey

Related Publication [↕](#)
 Bisgaard, Martin, and Rune Slothuus. [date]. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. <http://ajps.org/>

Notes [↕](#)
 This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. [Learn more about the Open Practice Badges from the Center for Open Science.](#)

Files | **Metadata** | Terms | Versions

Search this dataset...

Filter by
 File Type: All - Access: All - [↕](#)

1 to 10 of 25 Files [Download -](#)

<input type="checkbox"/>		build_data.R R Syntax - 12.1 KB Published Jun 29, 2020 56 Downloads MD5: a94...597	↕		
<input type="checkbox"/>		codebook_ess.pdf Adobe PDF - 508.8 KB Published Jun 29, 2020 46 Downloads	↕		

Dataset metrics

Files

Metadata

Terms

Versions

Citation Metadata ^

Dataset Persistent ID ?

doi:10.7910/DV

Previous Dataset Persistent ID ?

hdl:1902.1/000

Publication Date ?

2009-03-05

Title ?

Early Head Sta

Other ID ?

00097

Author ?

Administration for Children and Families (U.S. Department of Health & Human Services)

Description ?

This study page contains cataloging and documentation files (only) related to the *Early Head Start* data archived in the Murray Research Archive Dataverse.

The purpose of this study was to assess the impact of early head start programs in response to the 1994 Head Start reauthorization which established a special initiative for services to families with infants and toddlers. The study was a program evaluation with 1500 families in Early Head Start programs and 1500 in a control group with no program participation.

Rich support for
metadata standards in
human and machine
readable formats.

Export Metadata ▾

Dublin Core

DDI

DataCite

DDI HTML Codebook

JSON

OAI_ORE

OpenAIRE

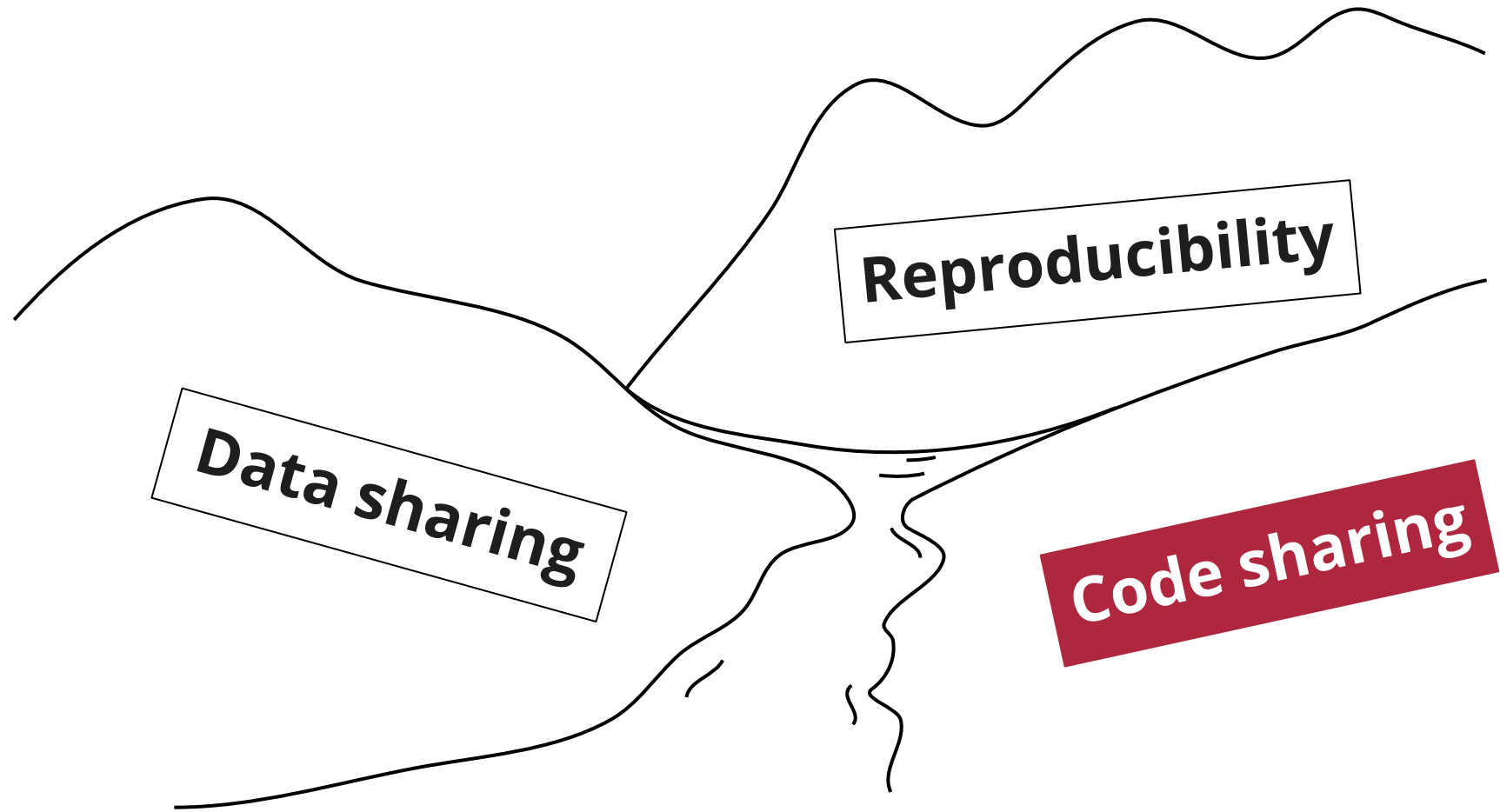
Schema.org JSON-LD



Data sharing

Summary

- FAIR principles are a standard for sharing research datasets
 - Data repositories implement FAIR with standard (and domain-specific) metadata, identifiers, data curation, etc.
-



Data sharing

Reproducibility

Code sharing

Research software includes code files, algorithms, scripts, workflows and executables created during the research process or for a research purpose.

- Research software is a critical component of research
 - Software intensive projects are a majority of publications
 - Most-cited papers are methods and software
 - Funders encourage open-source software
- Effort to recognize software are a valuable research output

Findable
Accessible
Interoperable
Reusable

Findable	Describe software with metadata, assign id Metadata are FAIR and searchable/indexable
Accessible	Retrievable by its identifier via standard protocol Metadata are accessible even when sw is not
Interoperable	Software interoperates with other software by exchanging data and/or metadata
Reusable	Software is usable (executable) and reusable (documented, licensed, extendable)



GitLab

- Long-term solution for sharing software?
- Where to store metadata?
- Software citation?

Citation File Format and `CITATION.cff`

```
cff-version: 1.2.0
message: "If you use this software, please cite it as below."
authors:
- family-names: "Lisa"
  given-names: "Mona"
  orcid: "https://orcid.org/0000-0000-0000-0000"
- family-names: "Bot"
  given-names: "Hew"
  orcid: "https://orcid.org/0000-0000-0000-0000"
title: "My Research Software"
version: 2.0.4
doi: 10.5281/zenodo.1234
date-released: 2017-12-18
url: "https://github.com/github/linguist"
```

main Branches Tags

Go to file

Add file

Code

About

A Ruby library for manipulating CITATION.cff files.

- yaml metadata
- sustainability attribution
- citation standard credit
- research-software-engineering

Readme

Apache-2.0 License

Cite this repository

Cite this repository

If you use this software in your work, please cite it using the following metadata. [Learn more](#)

APA BibTeX

Haines R. (2018). Ruby CFF Library (versio

View citation file

Commit	Message	Time
hainesr	Fix some minor issues with CFF fixtures. ...	11 days ago
db84460		11 days ago
288	commits	
.github/workflows	Turn Coveralls reporting back on after move to Actions.	25 days ago
bin	Turn on and fix rubocop Style/FrozenStringLiteralCom...	3 years ago
lib	Reference::new can now accept a block.	27 days ago
test	Fix some minor issues with CFF fixtures.	11 days ago
.gitignore	Remove the .ruby-* files from the repo.	last month
.rubocop.yml	Add CFF::File.open which accepts a block.	27 days ago
.rubocop_todo.yml	Reference::new can now accept a block.	
.simplecov	Turn Coveralls reporting back on after move to Actions	
CHANGES.md	Update CHANGES.md and CITATION.cff for release.	
CITATION.cff	Update the CITATION.cff file to add a comment.	
CODE_OF_CONDUCT.md	Add a code of conduct.	
Gemfile	Turn on and fix rubocop Style/FrozenStringLiteralCom	
LICENCE	Update the LICENCE and the file headers.	
README.md	Update README with new Model and File APIs.	

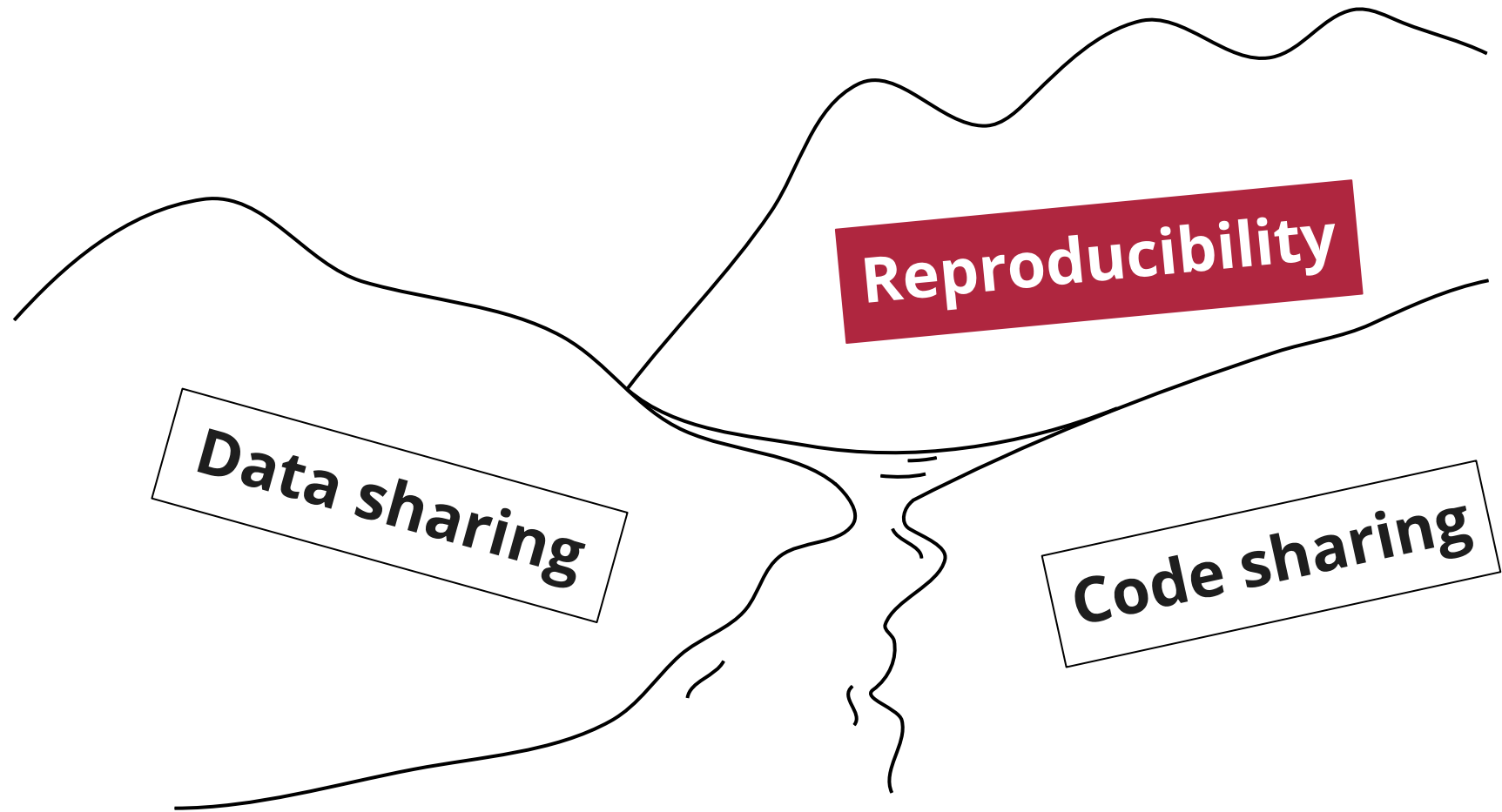
Software archival

- SoftwareHeritage.org
 - Initiative to collect, preserve and share software source code for the long-term future
- Zenodo
 - General purpose repository that support software deposits

Code sharing

Summary

- Research software as a recognized scientific output
 - FAIR principles for software help discoverability, licencing, attribution, citation, documentation, etc.
-



Reproducibility

Data sharing

Code sharing

Science has been in a “replication crisis” for a decade. Have we learned anything?

Bad papers are still published. But some other things might be getting better.

By Kelsey Piper | Oct 14, 2020, 12:20pm EDT

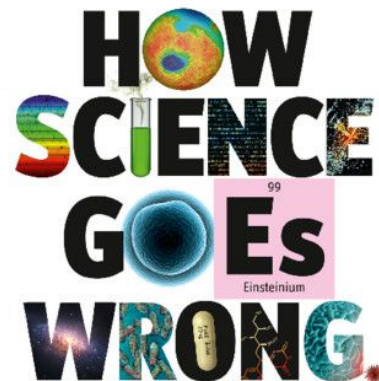
Home / News & Opinion

Potential Causes of Irreproducibility Revealed

Five independent groups got different results in a drug-response experiment, despite sharing protocols, reagents, and cell lines. The researchers identify technical variables could be to blame.

The Economist
OCTOBER 19TH-25TH 2018 Economist.com

- Washington's lawyer surplus
- How to do a nuclear deal with Iran
- Investment tips from Nobel economists
- Junk bonds are back
- The meaning of Sachin Tendulkar



nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [news feature](#) > [article](#)

[Published: 25 May 2016](#)

1,500 scientists lift the lid on reproducibility

[Monya Baker](#)

[Nature](#) 533, 452–454 (2016) | [Cite this article](#)

34k Accesses | 1489 Citations | 3920 Altmetric | [Metrics](#)

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science



National Academies of Sciences, Engineering,
and Medicine. 2019. Reproducibility and
Replicability in Science.

- Replication dataset - a bundle of data, code and other files needed to reproduce a published study

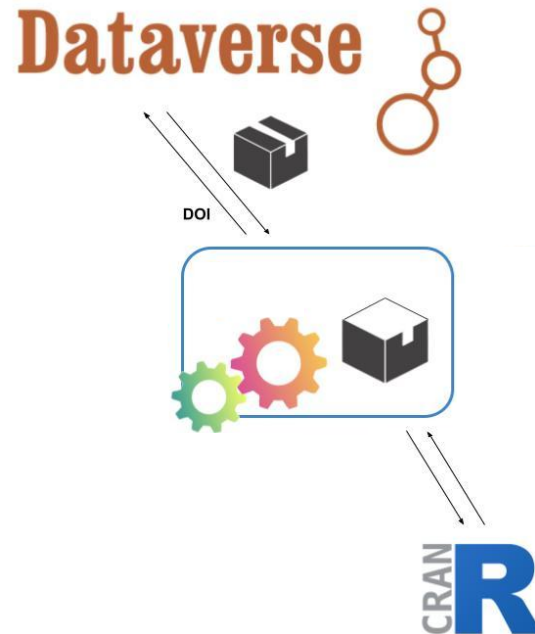
- Replication dataset - a bundle of data, code and other files needed to reproduce a published study

The screenshot shows the Harvard Dataverse interface. At the top, it says "HARVARD Dataverse" with navigation links for "Add Data", "Search", "About", "User Guide", and "Support". Below this is a large green banner with "AJPS" in large letters and "AMERICAN JOURNAL of POLITICAL SCIENCE" in smaller text. Underneath the banner, it reads "American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org". A breadcrumb trail shows "Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse". At the bottom, there is a paragraph stating: "The American Journal of Political Science is committed to significant advances in knowledge and understanding of citizenship, gov politics, and to the public value of political science research. To find out more about our data integrity policies, please visit our [web](#)."

The screenshot shows the detailed page for a replication dataset. At the top, it says "American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org". Below this is a breadcrumb trail: "Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse >". The main title is "Replication Data for: How Political Parties Shape Public Opinion in the Real World". A "Version 2.0" badge is visible. There is a "Cite Dataset" button and a link to "Learn about Data Citation Standards". A "Description" section explains the dataset's purpose and methodology. A "Subject" section lists "Social Sciences". A "Keyword" section lists "Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey". A "Related Publication" section lists a paper by Bisgaard, Martin, and Rune Slothuus. A "Notes" section mentions an independent verification process. There are also "Open Data" and "Open Materials" badges. At the bottom, there is a "Files" section with a search bar and a list of files: "build_data.R" (12.1 KB, 56 Downloads) and "codebook ess.pdf" (508.8 KB, 46 Downloads). Each file has a download icon and a view icon.

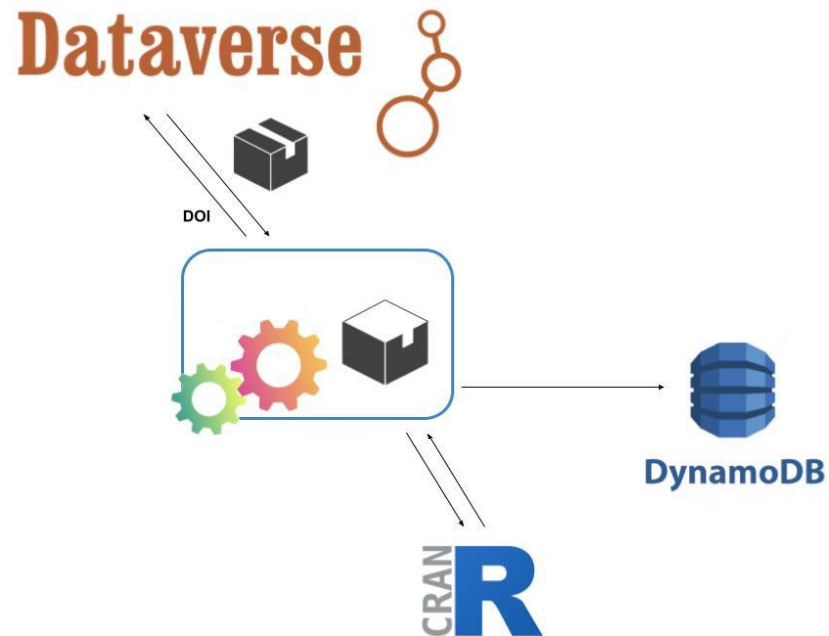
Our data collection workflow

1. Replication dataset is retrieved from Harvard Dataverse
2. We collect data on the content, install used R libraries and attempt automatic code re-execution

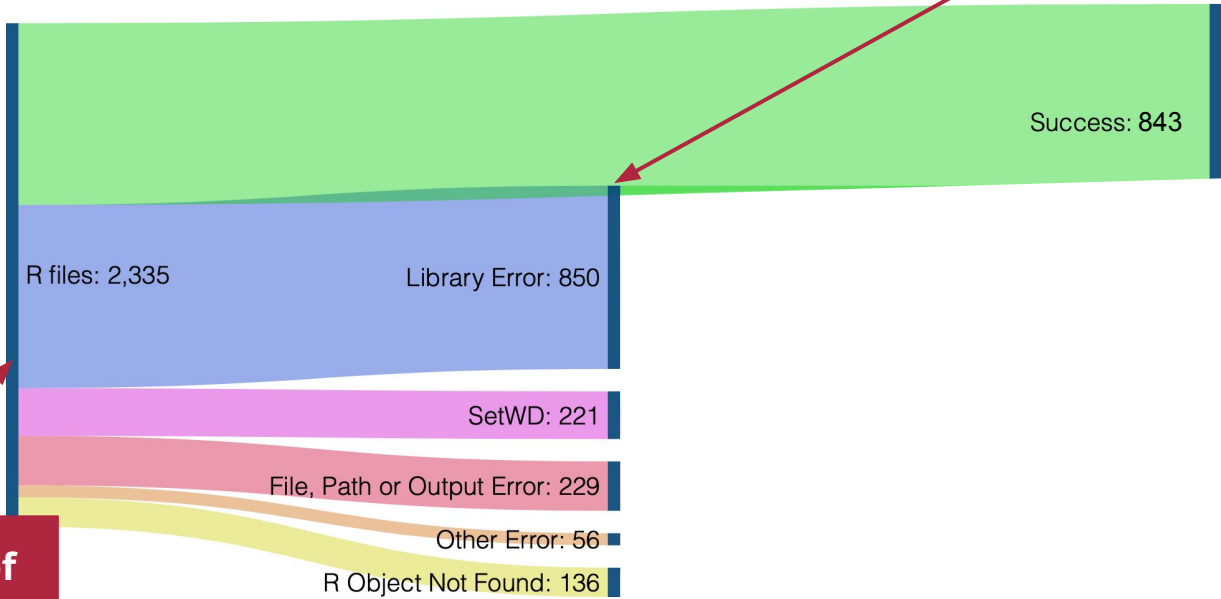


Our data collection workflow

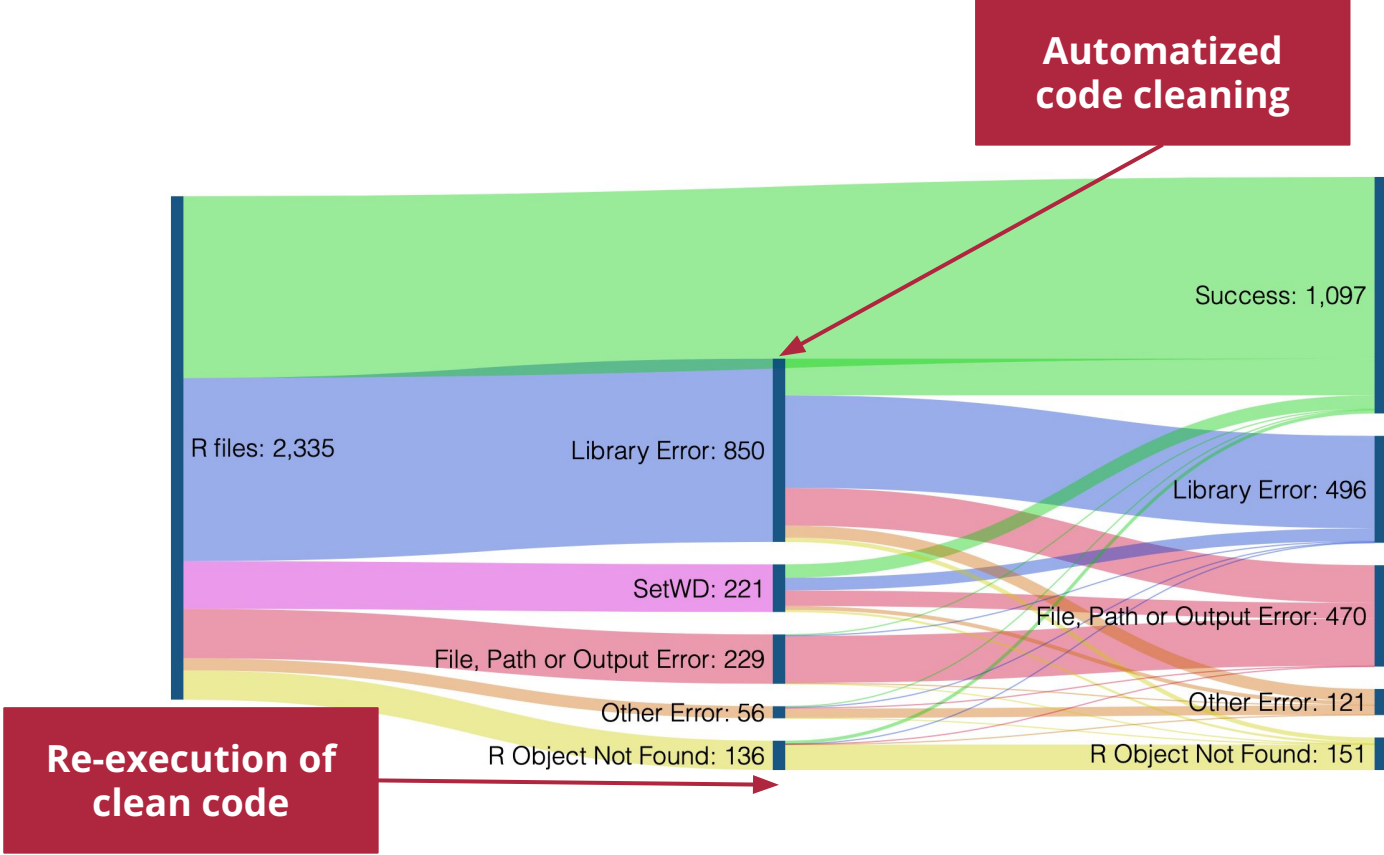
1. Replication dataset is retrieved from Harvard Dataverse
2. We collect data on the content, install used R libraries and attempt automatic code re-execution
3. The re-execution result and other collected data are passed to a database for analysis



Re-execution of original code

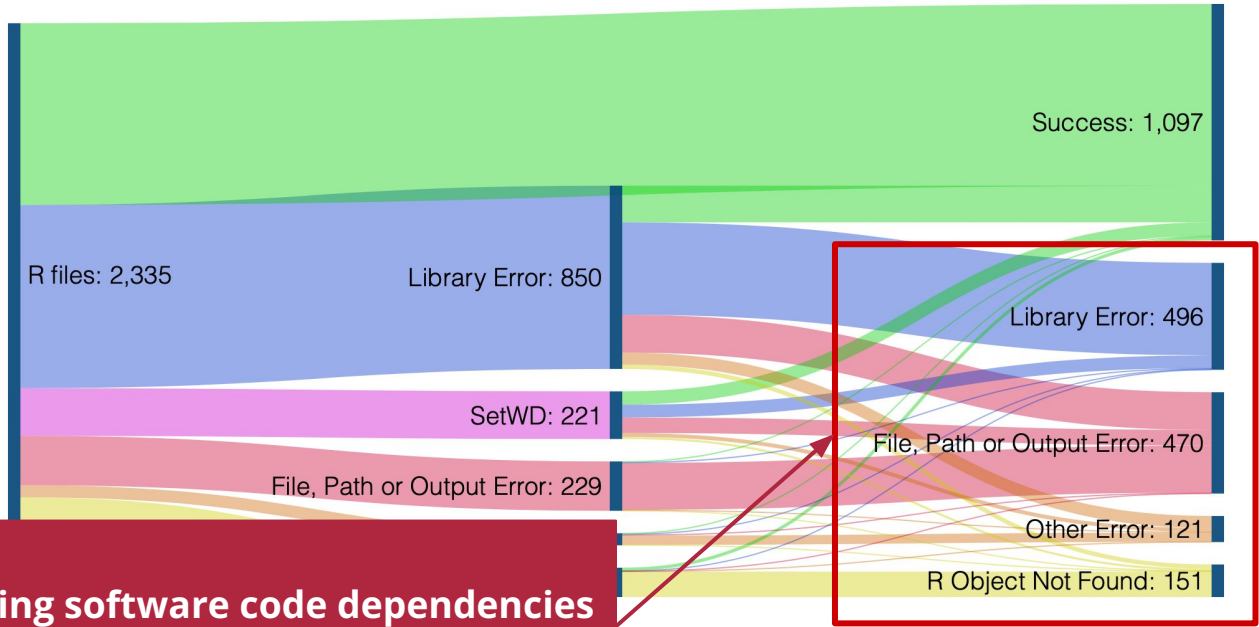


Automatized code cleaning

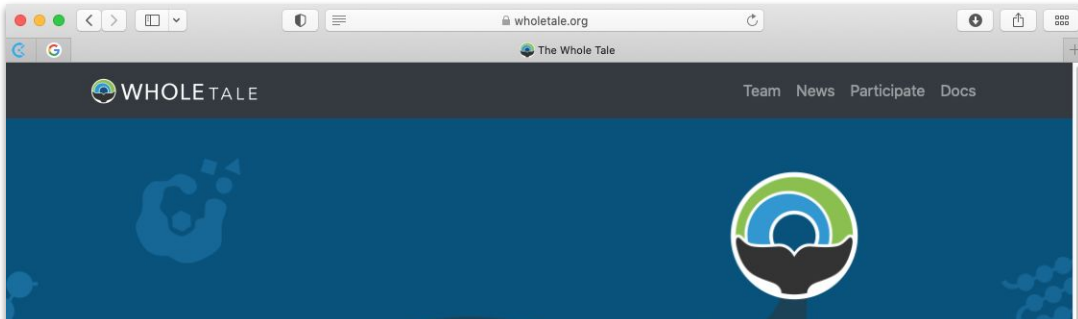


Automatized code cleaning

Re-execution of clean code



Capturing software code dependencies is critical for code sharing



The image shows the 'WHOLE TALE DASHBOARD' interface. The top navigation bar includes 'WHOLE TALE DASHBOARD', 'BROWSE', 'RUN', 'MANAGE', and 'COMPOSE', along with a user profile for 'Ana Trisovic'. The main content is divided into two panels. The left panel, titled 'Browse Tales', has a search bar and a filter set to 'All'. It displays three scientific tales: 'LIGO Tutorial LIGO Detected' with a LIGO waveform, 'SCIENCE Informatics-aided bandgap engineering...' with a scatter plot of Measured Band Gap (eV) vs. an unlabeled variable, and 'SCIENCE Accelerated discovery of metallic g...' with a diagram showing 'learning dels' leading to 'Predictions' and 'High-thr experimen'. The right panel, titled 'Launched Tales', shows a single tale: 'Predicting the Properties of Inorga...' with a Jupyter logo. The scatter plot in the 'Browse Tales' panel includes a legend for OLS (black dot), LASSO (green dot), and PLS (red dot), and a blue regression line.

Anyone can explore data, re-run code or modify it

Analysis of Python results

```
In [1]: import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import rc
import pandas as pd
# plot styles
sns.set_style('whitegrid')
sns.set_style({'font.family': 'Times New Roman'})
```

```
In [2]: df = pd.read_csv("python-study-data.csv", index_col=0)
```

```
In [3]: df.head()
```

Out [3]:

	doi	filename	result2	result3	list_of_all	size
0	doi:10.7910/DVN/8TB7GO	ei_preprocessing.py	TypeError: coercing to Unicode: need string or...	TypeError: invalid file: None	simulation_output.txt;format_cdc_data.sh;evalu...	274991
1	doi:10.7910/DVN/8TB7GO	ei_preprocessing_india.py	TypeError: unsupported operand type(s) for +: ...	SyntaxError: Missing parentheses in call to 'p...	simulation_output.txt;format_cdc_data.sh;evalu...	274991
2	doi:10.7910/DVN/8TB7GO	ei_preprocessing_ipums_census_acs_samples.py	AttributeError: 'NoneType' object has no attri...	SyntaxError: Missing parentheses in call to 'p...	simulation_output.txt;format_cdc_data.sh;evalu...	274991
3	doi:10.7910/DVN/8TB7GO	ei_preprocessing_ipums_full_census.py	AttributeError: 'NoneType' object has no attri...	SyntaxError: Missing parentheses in call to 'p...	simulation_output.txt;format_cdc_data.sh;evalu...	274991
4	doi:10.7910/DVN/8TB7GO	ei_preprocessing_race.py	TypeError: coercing to Unicode: need string or...	TypeError: invalid file: None	simulation_output.txt;format_cdc_data.sh;evalu...	274991

```
In [91]: df[df.result2.isnull()]
```

Out [91]:

	doi	filename	result2	result3	list_of_all	size
--	-----	----------	---------	---------	-------------	------



<https://www.reprozip.org>

```
vagrant@ubuntu-1604-amd64: ~/reprozip-examples/brain-segmentation 116x36
vagrant@ubuntu-1604-amd64:~/reprozip-examples/brain-segmentation$ reprozip trace python brain-segmentation.py
Configuration file written in .reprozip-trace/config.yml
Edit that file then run the packer -- use 'reprozip pack -h' for help
vagrant@ubuntu-1604-amd64:~/reprozip-examples/brain-segmentation$ eog median_otsu.png
```



ReproServer Upload

Select a package to upload

Upload a file

no file selected

or provide a package's URL

<https://www.reprozip.org>

```
vagrant@ubuntu-1604-amd64: ~/reprozip-examples/brain-segmentation 116x36
vagrant@ubuntu-1604-amd64:~/reprozip-examples/brain-segmentation$ reprozip trace python brain-segmentation.py
Configuration file written in .reprozip-trace/config.yml
Edit that file then run the packer -- use 'reprozip pack -h' for help
vagrant@ubuntu-1604-amd64:~/reprozip-examples/brain-segmentation$ eog median_otsu.png
```



CODE OCEAN



Stencila



colab

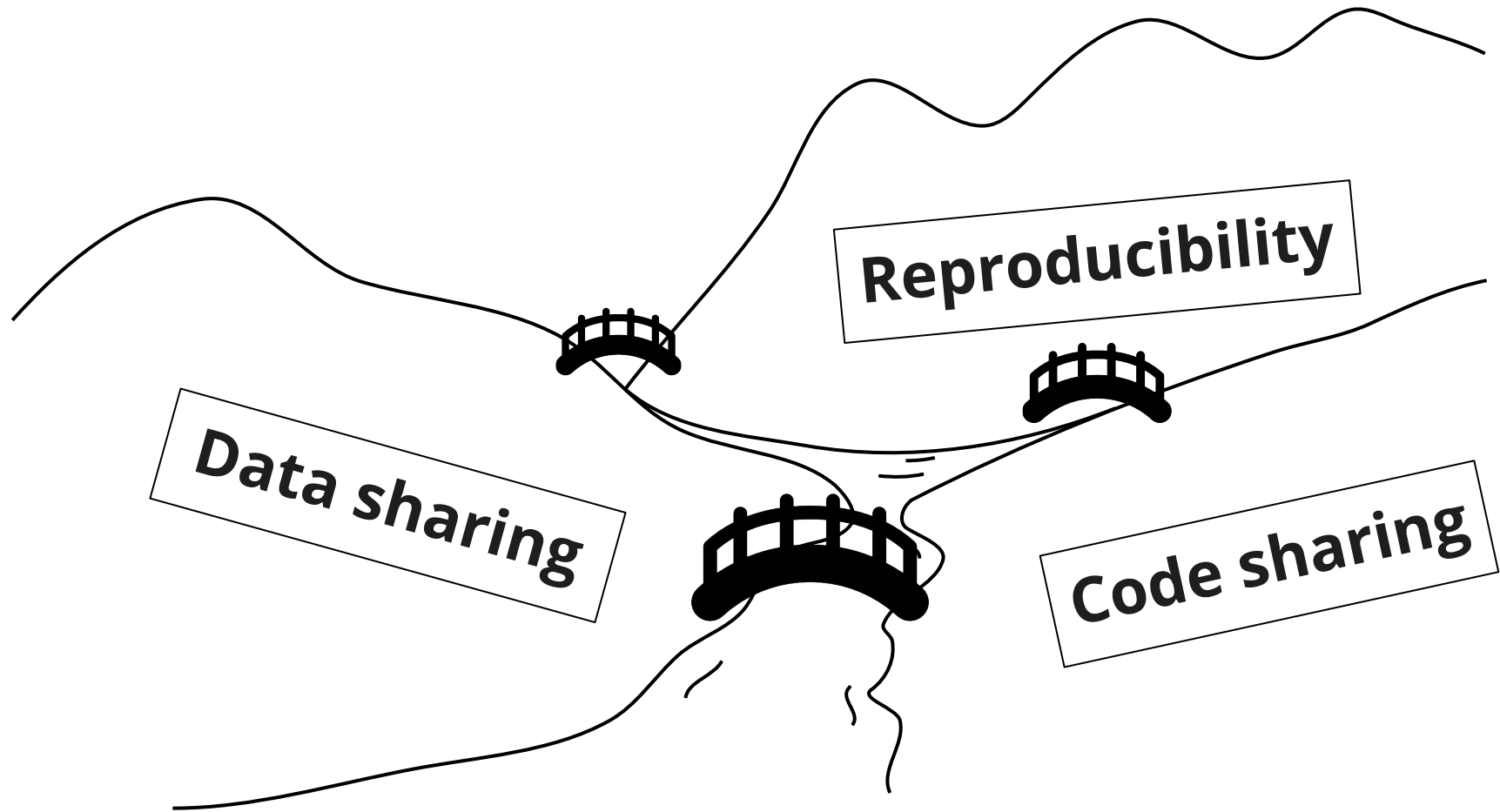


A hand-drawn diagram consisting of a central point from which three wavy lines branch out. Each branch terminates in a rectangular box containing text. The boxes are tilted to follow the direction of their respective branches. The text in the boxes is 'Data sharing', 'Reproducibility', and 'Code sharing'.

Reproducibility

Data sharing

Code sharing



Data sharing

Reproducibility

Code sharing

Data repository



Cloud platform

Data repository



Cloud platform

HARVARD
Dataverse

Add Data Search About User Guide Support Ana Trisovic

Replication Data for: Repository approaches to improving quality of shared data and code

Version 4.1

Trisovic, Ana, 2020, "Replication Data for: Repository approaches to improving quality of shared data and code", <https://doi.org/10.7910/DAN/EA3LC5>, Harvard Dataverse, V4

Cite Dataset Learn about Data Citation Standards.

Description This is supplementary data to the article "Repository approaches to improving quality of shared data and code," and in particular, its first section on completeness of research code.
Run this code on Jupyter Binder here: [launch binder](#) (2020-09-27)

Subject Computer and Information Science

Files Metadata Terms Versions

Search this dataset... Find Upload Files



Trisovic, Ana, et al. "Advancing computational reproducibility in the Dataverse data repository platform." P-RECS'20.

Data repository

Code repository



Cloud platform

Data repository

Code repository



Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

New to Binder? Get started with a Zero-to-Binder tutorial in [Julia](#), [Python](#) or [R](#).

Build and launch a repository

GitHub repository name or URL

GitHub repository name or URL

- GitHub
- Gist
- Git repository
- GitLab.com
- Zenodo DOI
- Figshare DOI
- Hydroshare resource
- Dataverse DOI

Path to a notebook file (optional)

Path to a notebook file (optional)

File

launch

and share your Binder with others:

to see a URL for sharing your Binder.

below paste it into your README to show a binder badge: [launch binder](#)



GitHub Action

Dataverse Uploader Action

v1.0 Latest version

[Use latest version](#)

Dataverse Uploader

This action uploads the repository content to a Dataverse dataset.

Input parameters

To use this action, you will need the following input parameters:

Parameter	Required	Description
<code>DATVERSE_TOKEN</code>	Yes	This is your personal acc token that you can create your Dataverse instance (the Dataverse guide). Sa your token as a secret variable called <code>DATVERSE_TOKEN</code> in yo

Stars: [☆ Star](#) 0

Contributors:

Categories: [Publishing](#) [Utilities](#)

Links: [atrisovic/dataverse-uploader](#) [Open issues](#) 0 [Pull requests](#) 0 [Report abuse](#)



Software Heritage

GitHub Action

Save to Software Heritage

v1.0.1 Latest version

Use latest version

Software Heritage Save action

A GitHub Action that saves the GitHub repository it is being run on to the [Software Heritage Archive](#).

Inputs

n/a - Action can only save repository that it is run on. Also prevents misuse.

Outputs

result

The result string from the call to the Software Heritage API. To track the actual save result, go to <https://archive.softwareheritage.org/save/#requests> and look for the name of your repository.

Stars

☆ Star 6

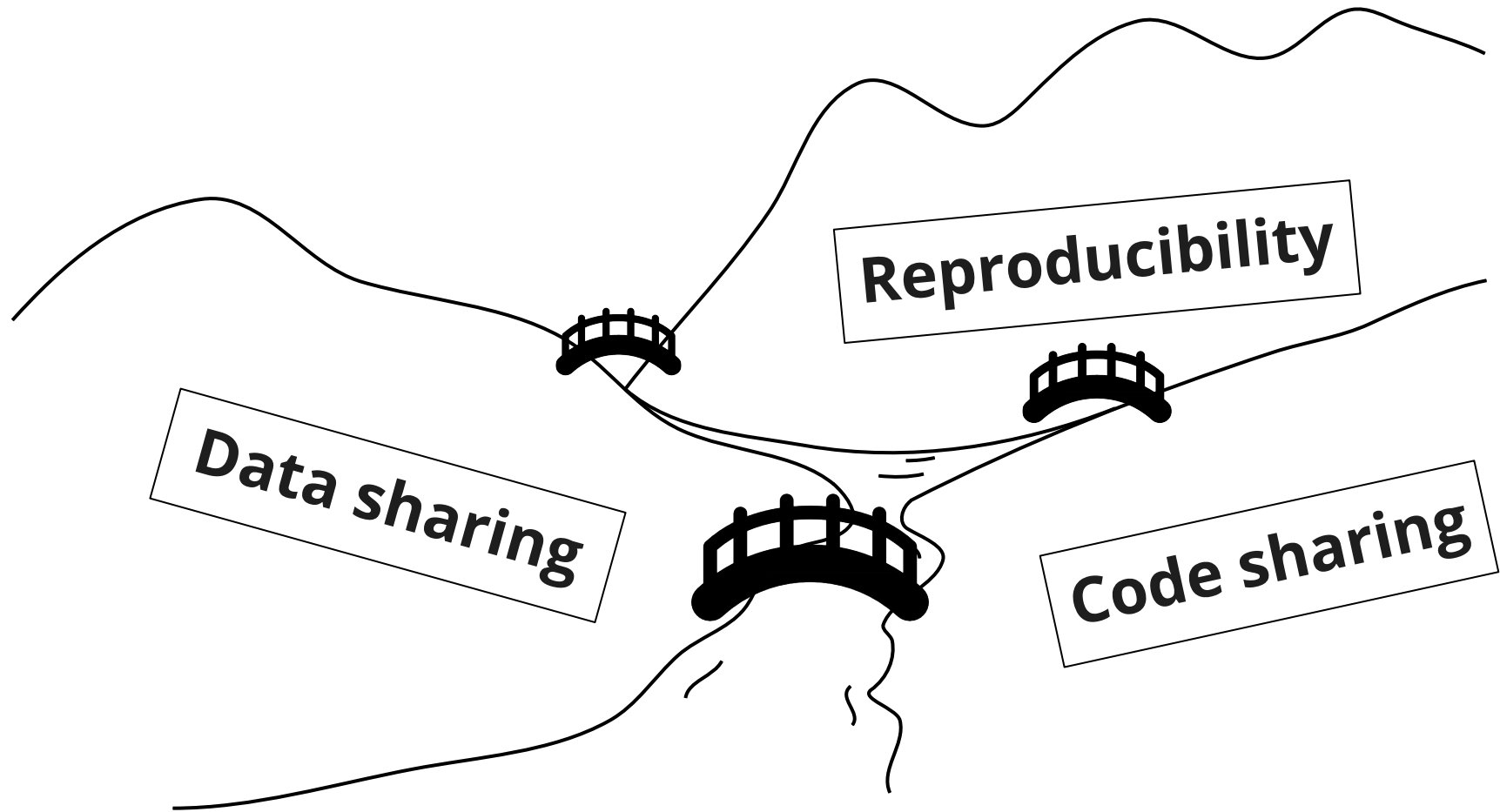
Contributors

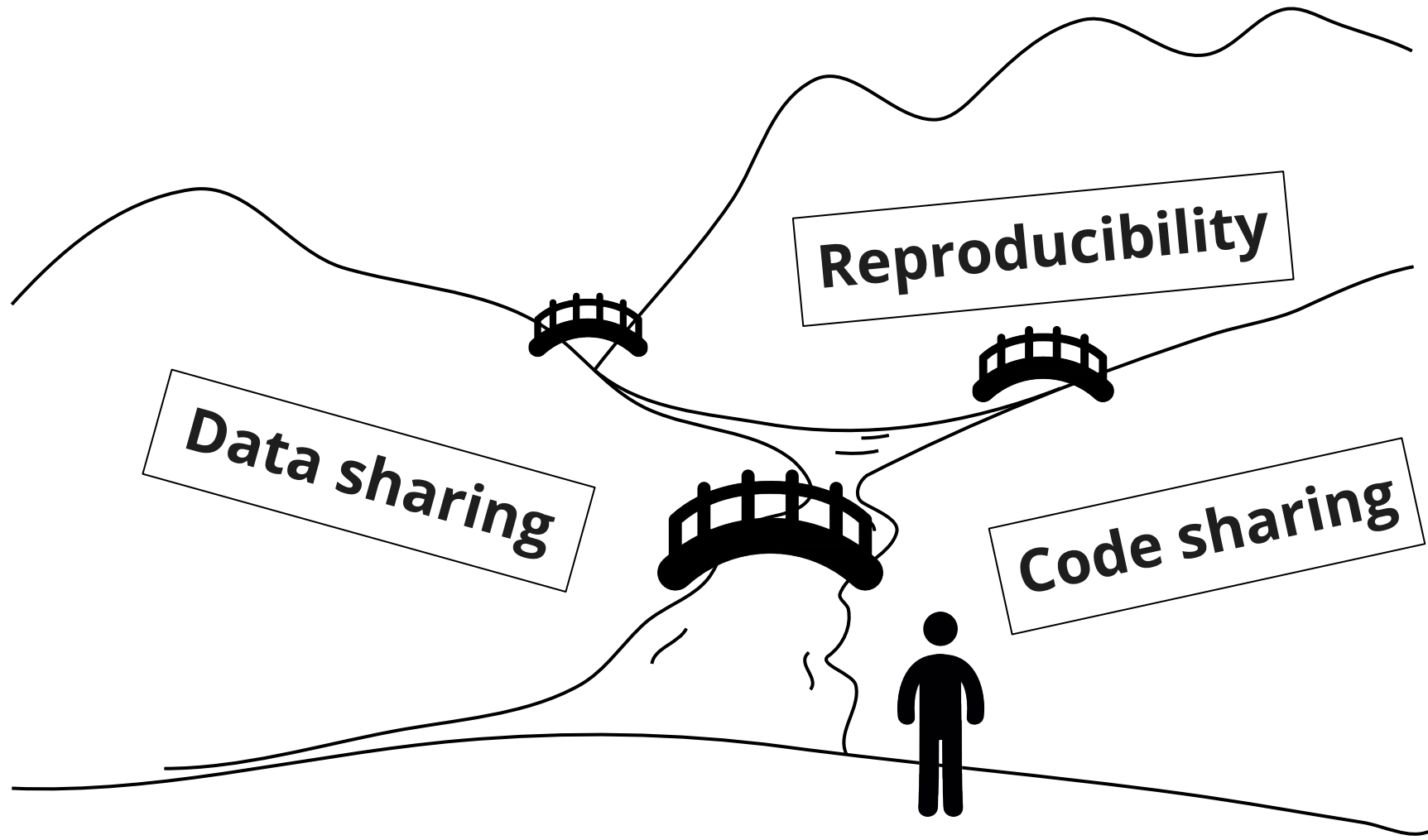
Categories

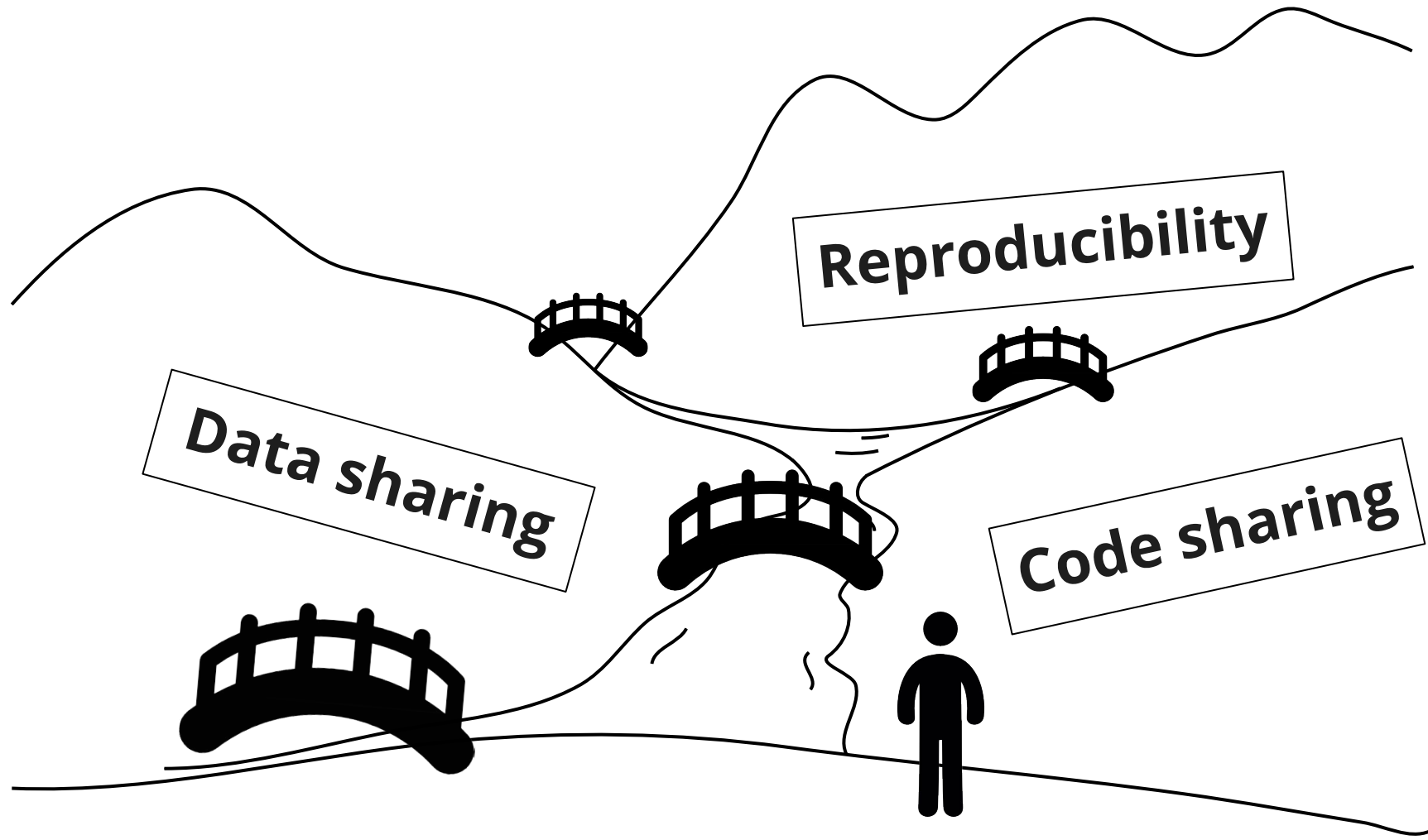
[Open Source management](#) [Backup Utilities](#)

Links

- [sdruskat/swh-save-action](#)
- [Open issues](#) 0
- [Pull requests](#) 0
- [Report abuse](#)







1. Data sharing

1. Data sharing

HARVARD
Dataverse

Add Data Search About User Guide Support Sign Up Log In

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

[NSAPH Dataverse](#) (Harvard University)

[Harvard Dataverse](#) >

[Contact](#) [Share](#)

Welcome to the dataverse collection by the National Studies on Air Pollution and Health group at the Harvard T.H. Chan School of Public Health. The group releases analysis and open data relating to air quality, demographics, and health.

Search this dataverse... [Advanced Search](#)

[Dataverses \(2\)](#)
 [Datasets \(2\)](#)
 [Files \(18\)](#)

Dataverse Category
[Research Group \(2\)](#)


Publication Year
[2022 \(4\)](#)

Subject
[Earth and Environmental Sciences \(1\)](#)
[Mathematical Sciences \(1\)](#)
[Medicine, Health and Life Sciences \(1\)](#)

Author Name
[Braun, Danielle \(1\)](#)
[Khoshnevis, Naeem \(1\)](#)
[Woodward, Sophie \(1\)](#)
[Wu, Xiao \(1\)](#)

1 to 4 of 4 Results [Sort](#)

Replication Data for: Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States
Feb 14, 2022 - NSAPH Analysis Data

 Woodward, Sophie, 2022, "Replication Data for: Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States", <https://doi.org/10.7910/DVN/3ZU0AS>, Harvard Dataverse, V1, UNF:6:D8W7/RcF4rlkKWjYBggFeQ== [fileUNF]

This is the data repository for publicly available data to reproduce analyses in Woodward, S., Wu, X., Hou, Z., Mork, D., Braun, D., Dominici, F., 2022. Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mort...

NSAPH Open Data (Harvard University)
Feb 7, 2022

NSAPH Analysis Data (Harvard University)
Feb 7, 2022

1. Data sharing

HARVARD
Dataverse

Add Data ▾ Search ▾ About User Guide Support Sign Up Log In

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NSAPH Dataverse (Harvard University)

Harvard Dataverse >

Contact Share

Welcome to the dataverse collection by the National Studies on Air Pollution and Health group at the Harvard T.H. Chan School of Public Health. The group releases analysis and open data relating to air quality, demographics, and health.

Search this dataverse... [Advanced Search](#)

Dataverses (2)
 Datasets (2)
 Files (18)

Dataverse Category
[Research Group \(2\)](#)

Publication Year
[2022 \(4\)](#)

Subject
[Earth and Environmental Sciences \(1\)](#)
[Mathematical Sciences \(1\)](#)
[Medicine, Health and Life Sciences \(1\)](#)

Author Name
[Braun, Danielle \(1\)](#)
[Khoshnevis, Naeem \(1\)](#)
[Woodward, Sophie \(1\)](#)
[Wu, Xiao \(1\)](#)

1 to 4 of 4 Results

Replication Data for: Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States
Feb 14, 2022 - NSAPH Analysis Data

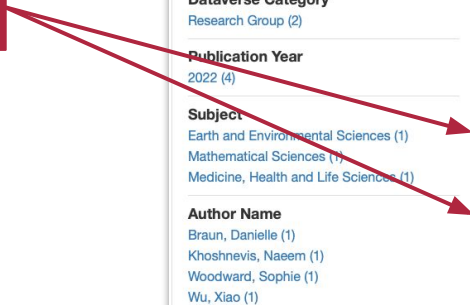
Woodward, Sophie, 2022, "Replication Data for: Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States", <https://doi.org/10.7910/DVN/3ZU0AS>, Harvard Dataverse, V1, UNF:6:D8W7/RcF4rlkKWjYBggFeQ== [fileUNF]

This is the data repository for publicly available data to reproduce analyses in Woodward, S., Wu, X., Hou, Z., Mork, D., Braun, D., Dominici, F., 2022. Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mort...

NSAPH Open Data (Harvard University)
Feb 7, 2022

NSAPH Analysis Data (Harvard University)
Feb 7, 2022

Data collections










Data that does not explicitly have an open license is not open data.

1. Data sharing

- Data should be licensed
- Metadata
- It should be complete
- It should be shared in a (open) machine-readable format

1. Data sharing

- Data should be licensed
- Metadata
- It should be complete
- It should be shared in a (open) machine-readable format

Name	Date Modified	Size	Kind
 ATP W85 methodology.pdf	Nov 11, 2021 at 4:13 PM	138 KB	PDF Document
 ATP W85 questionnaire.pdf	Nov 12, 2021 at 12:51 PM	1.7 MB	PDF Document
 ATP W85 readme.txt	Jan 26, 2022 at 1:04 PM	4 KB	Plain Text
 ATP W85 topline.pdf	Nov 11, 2021 at 4:17 PM	294 KB	PDF Document
 ATP W85.sav	Jan 26, 2022 at 12:43 PM	2.1 MB	Document



2. Data documentation

Data Dictionary

Fieldname	Source	Description	Role
QID	Medicare	Person's ID	ID
ADATE	Medicare	Admission date	Outcome
DDATE	Medicare	Discharge date	Outcome
zipcode_R	Medicare	Zipcode	Location
DIAG 1-10	Medicare	Billing codes as ICD codes	Outcome
AGE	Medicare	Age	Confounder
PROV_NUM	Medicare	Hospital ID	Location
ADM_SOURCE	Medicare	Admission source	Confounder
ADM_TYPE	Medicare	Admission type (1 - Emergency, 2 - Urgent, 3 - Elective)	Confounder
Dual	Medicare	Eligible for both Medicare and Medicaid (1 - yes, 0 - otherwise)	Confounder
year	R	Year of hospital admission (from 2000 to 2016)	Outcome
AD_primary	R	Does the ICD code of Alzheimer's disease appear in the first billing code? (T/F)	Outcome

2. Data documentation

Data Dictionary

Fieldname	Source	Description	Role
QID	Medicare	Person's ID	ID
ADATE	Medicare	Admission date	Outcome
DDATE	Medicare	Discharge date	Outcome
zipcode_R	Medicare	Zipcode	Location
DIAG 1-10	Medicare	Billing codes as ICD codes	Outcome
AGE	Medicare	Age	Confounder
PROV_NUM	Medicare	Hospital ID	Location
ADM_SOURCE	Medicare	Admission source	Confounder
ADM_TYPE	Medicare	Admission type (1 - Emergency, 2 - Urgent, 3 - Elective)	Confounder
Dual	Medicare	Eligible for both Medicare and Medicaid (1 - yes, 0 - otherwise)	Confounder
year	R	Year of hospital admission (from 2000 to 2016)	Outcome
AD_primary	R	Does the ICD code of Alzheimer's disease appear in the first billing code? (T/F)	Outcome

Variable Description

	ADM_SOURCE - Admission source
1	Physician referral
2	Clinic referral
3	HMO referral
4	Transfer from hospital
5	Transfer from a SNF
6	another health care facility
7	Emergency room

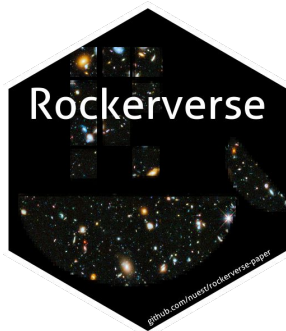
3. Survey data documentation

- What is the topic of your survey?
- What are the questions? What are the allowable answers?
- Special variables - such as a score?
- Survey administration
 - Who enters the data?
 - Paper survey (errors handling, interviews)
 - Online survey (skip patterns)

4. Use version control for code



5. Runtime environment capture



Pin dependencies

When you install a dependency, include its version number (depending on the language you use, the exact syntax may vary). E.g., don't just specify `numpy`, specify `numpy==1.12.0`.

`pip freeze` is a handy tool to export the exact version of every Python package in your environment in a format that can be used in `requirements.txt`.

`conda env export -n <env-name>` is the equivalent for anaconda's `environment.yml` file.

When exporting a conda environment, you can add the conda-forge broken channel (`conda-forge/label/broken`) as a low-priority channel in your exported `environment.yml` file in order to maximize durability. Thus, if a package is marked broken after you froze the environment, said package will still install during the Binder image build process. Only do this when you intend to truly freeze the environment.

For example (in `environment.yml`):

```
channels:  
- conda-forge  
- defaults  
- conda-forge/label/broken
```

Using Dockerfiles

Ensuring reproducibility with Dockerfiles comes with its own set of challenges. For more

<https://mybinder.readthedocs.io/en/latest/tutorials/reproducibility.html>


6. Automation



COMMON
WORKFLOW
LANGUAGE

7. Internal review

add data dictionary in README #1

 Merged ShuxinD merged 5 commits into `ShuxinD:main` from `atrisovic:patch-1`  2 days ago

 Conversation 7  Commits 5  Checks 0  Files changed 1



atrisovic commented 8 days ago

Contributor  ...

No description provided.



atrisovic added 2 commits 8 days ago



Update README.md

Verified e2d8f5a



Update README.md

Verified 9bc59c8



ShuxinD reviewed 8 days ago

View changes

README.md Outdated

Show resolved



ShuxinD reviewed 8 days ago

View changes

README.md Outdated

Hide resolved

```
29 + | ADATE | Medicare | Admission date | Time |
30 + | DDATE | Medicare | Discharge date | Time |
31 + | zipcode_R | Medicare | Zipcode | Location |
32 + | DIAG 1-10 | Medicare | Diagnosis code | Outcome |
```



ShuxinD 8 days ago

Owner  ...

How about changing "Diagnosis code" to "Billing codes as ICD codes"

Recommendations

Summary

- Use data repositories for data sharing
 - Document your data
 - Capture dependencies of your code
 - Establish internal review workflow
-

Email: anatrisonic@g.harvard.edu
GitHub & Twitter: atrisonic

